Acta Scientiarum

http://periodicos.uem.br/ojs
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v45i1.60184

STATISTICS

Feasibility of the expectation-maximization algorithm for assessing individuals with different sensory perceptions in discrimination of specialty coffees

Larissa Karolina de Oliveira[1], Mariana Resende[2], Flávio Meira Borém[3] and Marcelo Angelo Cirillo[2*]

[1]Departamento de Engenharia, Universidade Federal de Lavras, Câmpus Universitário, Cx. Postal 3037, 37200-900, Lavras, Minas Gerais, Brazil. [2]Departamento de Estatística, Universidade Federal de Lavras, Lavras, Minas Gerais, Brazil. [3]Departamento de Engenharia Agrícola, Universidade Federal de Lavras, Lavras, Minas Gerais, Brazil. *Author for correspondence. E-mail: macufla@gmail.com

ABSTRACT. The results of sensory evaluations of coffees are associated with latent factors, such as the particular subjectivity of each individual. Based on the foregoing, assessing the quality of a sensory panel for product discrimination basically depends on the statistical methodology to be used in data analysis. Following this argument, this study aimed to evaluate the feasibility of the EM - Expectation Maximization algorithm in discriminating groups of individuals, characterized by the degree of experience and knowledge in sensory analysis of coffees of different varieties, produced in the Serra da Mantiqueira micro-region, with different processing and altitudes. The main advantage of this algorithm is the fast convergence, when the current solution approaches the optimal solution with high precision. The disadvantage is because it is a deterministic optimization technique, which can only achieve a local optimization depending on the initialization, i.e., initial values input in the iterative procedure. It can be concluded that estimates of the correlation matrices obtained by the EM algorithm showed that the final grade has a greater influence of sweetness, in addition to discriminating groups of consumers with different sensory perceptions and in situations where the number of individuals in each group is unknown, the EM algorithm was accurate in estimating the proportion of individuals belonging to each group, assuming that the correlations of sensory responses follow a bivariate normal distribution.

Keywords: mixture of distributions; latent variable; bivariate normal; body; acidity.

Received on July 19, 2021.
Accepted on April 14, 2022.

Introduction

According to the Specialty Coffee Association of America (SCAA, 2009), a specialty coffee is defined as a coffee of high quality, making the product unique, with particularities of the designation of origin and the sensory attributes, such as aroma, flavor, and in particular the selection of perfect grains listed and highlighted. All these factors contribute to the search for certification and the infiltration of new consumer niches.

The issue is that to identify and to determine the attributes that characterize the excellence of specialty coffee, it is necessary to conduct experiments involving sensory analysis. Thus, regardless of the quality of tasters, there is always the possibility that behavioral attitudes, or even the arrangement of samples to be tested, may influence the result of certain evaluations (Cirillo et al., 2019). This opens up a wide field of research on assessing the effect of the homogeneity of individuals in a sensory panel.

In the case of coffee quality analysis, this problem may become more critical if we consider a chemometric perspective, according to which the analysis of sensory attributes, such as the aroma, is based on olfactometry, centered on individual and consumer preferences (Yeretzian et al., 2010) due to reflexes and psychological perceptions that influence the discernment of a product's quality.

Other consumption factors may be contextualized in activities of individuals in a sensory analysis, involving different theories that explain different attitudes worthy of being identified in sensory experiments (Brookes, 2014; Sørensen, Andersen, & Andersen, 2014; Oliver, 2014; Ceribeli, 2015; Dai, Luo, Liao, & Cao, 2015; Ossani, Cirillo, Borem, Ribeiro, & Cortez, 2017). The main theories are discussed by Lankton and McKnight (2012) and include assimilation theory, which indicates that individuals tend to be reluctant to accept discrepancies from previously assumed positions.

Makarem and Jae (2016) mention that emotions perceived by the consumer and/or taster are identified as a set of hedonic responses that arise during the experiment, such as facial expressions, caused by a particular sample of coffee that has a higher degree of acidity or flavor and can cause reactions, such as pleasure or displeasure, calmness or excitement. Gashgari (2016) proposes the equity theory, which refers to the perception of the individual in convincing themselves of a certain result, as they observe a trend or acceptance by most tasters and/or consumers.

Regarding the perspectives of these theories, the selection of individuals to compose a sensory panel, as well as the discrimination of products, contextualized in this study by the different types of coffees, is complex because these psychological perceptions are not measurable. However, some models suggest considering these perceptions as latent variables for which predictions are made.

This motivates the proposal to use the expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), divided into two steps of step 'E' (expectation), which aims to calculate the expected value of the logarithm of the likelihood defined with all parameters, including the latent variable, and step 'M' (maximization), which finds the local maximum until reaching a numerical convergence defined by a margin of error.

The main advantage of this algorithm is the fast convergence, when the current solution approaches the optimal solution with high precision. The disadvantage is because it is a deterministic optimization technique, which can only achieve a local optimization depending on the initialization, i.e., initial values input in the iterative procedure (Li, Zhang, He, Tian, & Wei, 2019). Based on the above, in this study, this algorithm is applied to a sensory test, in which the results related to sensory analysis of four specialty coffees produced in the Serra da Mantiqueira region, with different genotypes, altitudes, and processing types, were obtained for two groups of tasters, trained and untrained individuals.

In this context, with combination of these samples and simulation of a situation in which the ability of tasters to discriminate the coffees is not known, the main objective of this study was to analyze the feasibility of estimating the proportion of individuals in each group and by comparing the estimate to the exact proportion based on the hypothesis that there is a correlation between each attribute with the overall score, supposedly with a bivariate normal distribution.

Material and methods

The methodology used in this study consisted of two stages, which are described in the following sections: Database description and feasibility of the EM algorithm for distinguishing groups in the sensory panel and Algebraic procedure used for implementation of the EM algorithm.

Database description and feasibility of the EM algorithm for distinguishing groups in the sensory panel

Results of analysis of the main sensory attributes, score, body, acidity, and sweetness were obtained for specialty coffees produced in the Serra da Mantiqueira region. Sensory experiments were carried out at the Federal University of Lavras. A preparation of the 100% arabica coffee samples was carried out by removing all defective and toasted grains and respecting the maximum period of 24 hours for taste testing.

The roasting point was visually determined using the color classification system by means of standardized disks (SCAA/AgtronRoast Color Classification System). With regard to beverage preparation, a concentration of 7% w v^{-1} was maintained using filtered water ready for consumption, free from contamination and without added sugar. Following these specifications, four types of specialty coffees coded as samples A, B, C, and D were used (Table 1). The evaluated specialty coffees produced in the Serra da Mantiqueira region differed in terms of processing, altitude, and genotype and are described in Table 1.

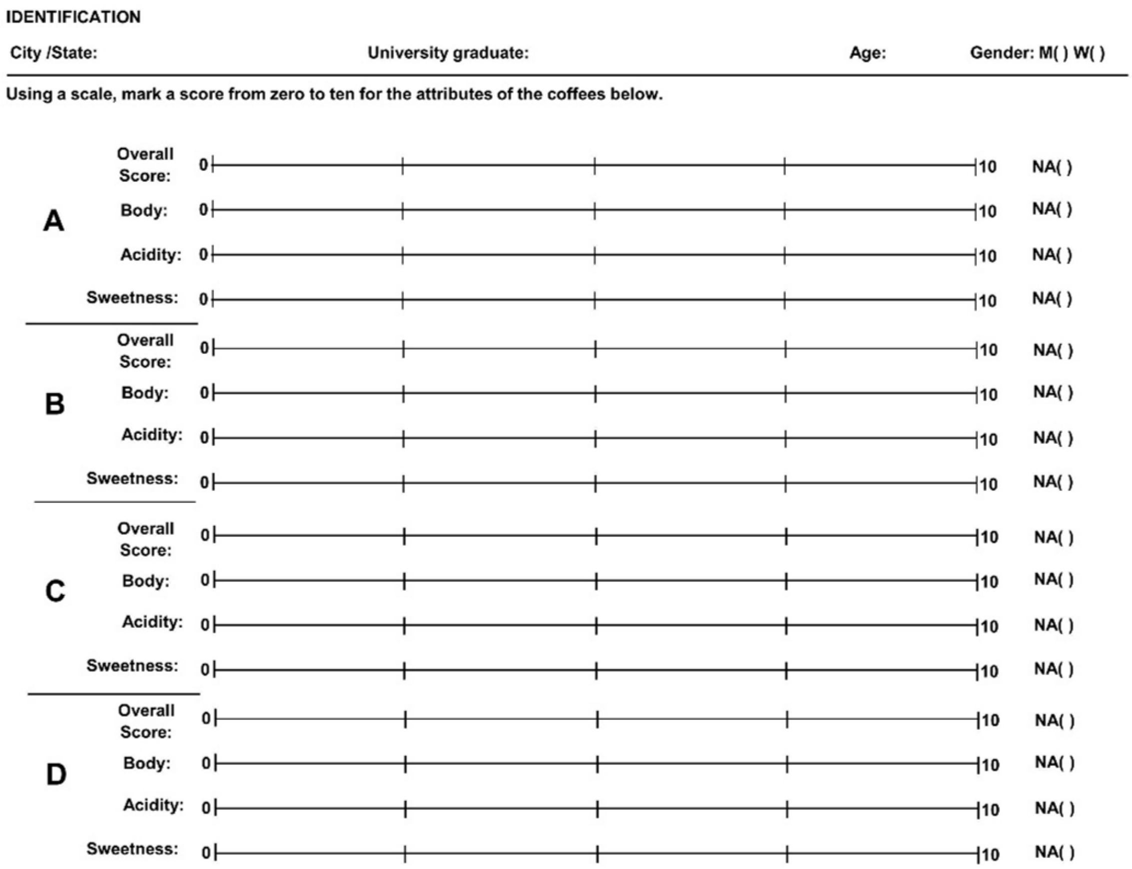
Table 1. Description of specialty coffees evaluated in the sensory analysis with trained and untrained consumers.

Type	Genotype	Altitude	Processing
A	Bourbon Amarelo	above 1,200 m	Natural
B	Acaiá	below 1,100 m	Pulped coffee cherry
C	Acaiá	below 1,100 m	Natural
D	Bourbon Amarelo	above 1,200 m	Pulped coffee cherry

Accordingly, sensory tests were carried out considering two distinct groups of consumers, where G1 consisted of consumers who received sensory evaluation training and G2 consisted of individuals who did not receive any training but were technicians or researchers in the field of coffee research. The numbers of individuals belonging to each group were $n_1 = 43$ and $n_2 = 57$, respectively. Therefore, the exact proportion of consumers in each group was known and given by $\phi_1 = 0.43$ and $\phi_2 = 0.57$.

In different sessions, each consumer evaluated the quality of each coffee on a continuous scale from 0 to 10 points for the following attributes: score, body, acidity, and sweetness, as shown in Figure 1.

For better clarification of the feasibility of using the EM algorithm for discrimination of sensory results for the different attributes (Figure 1), the estimates obtained in the execution of the EM algorithm, defined by $\widehat{\emptyset_j}$ (j = 1.2), were interpreted. Thus, given the proximity to the exact proportion, it became possible to judge the feasibility in discriminating the groups. A detailed description of the EM algorithm is provided below.


IDENTIFICATION
City /State: University graduate: Age: Gender: M( ) W( )
Using a scale, mark a score from zero to ten for the attributes of the coffees below.

A
Overall Score: 0 10 NA( )
Body: 0 10 NA( )
Acidity: 0 10 NA( )
Sweetness: 0 10 NA( )

B
Overall Score: 0 10 NA( )
Body: 0 10 NA( )
Acidity: 0 10 NA( )
Sweetness: 0 10 NA( )

C
Overall Score: 0 10 NA( )
Body: 0 10 NA( )
Acidity: 0 10 NA( )
Sweetness: 0 10 NA( )

D
Overall Score: 0 10 NA( )
Body: 0 10 NA( )
Acidity: 0 10 NA( )
Sweetness: 0 10 NA( )

Figure 1. Sensory form used in sensory analysis of specialty coffees described in Table 1.

Algebraic procedure used for implementation of the EM Algorithm

For execution of the EM algorithm, it was necessary to assign the following coding: for sensory attributes body, acidity, sweetness, and overall score, it was assumed that x_h (hour = 1...,4).

The type of coffee was identified following the description in Table 1 and was given by (k = 1..., 4). Based on the above, the vector of bivariate responses $\mathbf{x_k} = (x_h, x_1)$ (hour = 2..., 4) was considered for each consumer group G_j (j = 1 and 2), together with parameters represented in the bivariate covariance defined by $\boldsymbol{\Sigma_{jk}}$ with the respective vector of means $\boldsymbol{\mu_{jk}}$.

Following these specifications, given the continuous nature of the scale, a bivariate normal distribution was assumed. Thus, the vector of parameters to be estimated was set at $\theta = (\mu_{1k}, \mu_{2k}, \Sigma_{1k}, \Sigma_{2k}, \phi_{1k}, \phi_{2k})$, where ϕ_j (j = 1, 2) refers to the latent variable to be evaluated as the main parameter of interest to be estimated and compared with the exact proportion.

This validation enabled the assessment of the feasibility of using the EM algorithm in discriminating different groups of consumers for the case study herein. Next, a detailed description of the EM algorithm is given based on the expressions.

Step E of the multivariate EM algorithm adopted for bivariate normal distributions requires the likelihood function $L(\theta|x, z) = P(\theta|x, z)$ of the estimated parameter θ. The probability density function of the multivariate normal distribution is defined by (Equation 1).

$$f(x \vee \mu, \Sigma) = \frac{1}{2\pi\sqrt{\Sigma}\vee\exp\left(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}\quad(x-\mu)\right)} \tag{1}$$

The above probability L(θ|x, z) can be presented in separate sum terms using the indicator function (Equation 2):

$$L \text{ onde}, I_{jk}(Z_h) = \{\begin{matrix}1, sez_h = j \\ 0, sez_h \neq j\end{matrix} \tag{2}$$

Combining the probabilities in all observations (Equation 3):

$$L \tag{3}$$

Thus, we have the log of the likelihood required in step E of the multivariate algorithm (Equation 4):

$$logL\Sigma_{jk}^{-1}(x_h - \mu_{jk}) \tag{4}$$

Developing step E, where values of weights are assigned, we have (Equation 5):

$$P(z_h = j|x_h, \theta) = T_{j,h}^{(t)} = \frac{\phi_j^{(t)} f(x_h \vee \mu_{jk}^{(t)}, \Sigma_{jk}^{(t)}\quad)}{\phi_{1k}^{(t)} f(x_h \vee \mu_{1k}^{(t)}, \Sigma_{1k}^{(t)}\quad) + \phi_{2k}^{(t)} f(x_h \vee \mu_{2k}^{(t)}, \Sigma_{2k}^{(t)}\quad)} \tag{5}$$

where:
$T_{h,j}^{(t)}$ is the probability of the h-th sample belonging to the j-th distribution, interactively considering the previous estimate.

Thus, step E can be expressed as (Equation 6):

$$Q(\theta) = E\{logL(\theta|x,z)\} = E\{\sum_{h=1}^{N} logL(\theta|x_h, z_h)\} = \sum_{h=1}^{N} ElogL(\theta|x_h, z_h) \sum_{h=1}^{N}\sum_{jk=1}^{2} T_{jk,h}^{(t)} \tag{6}$$

In step M, to obtain updated parameter estimates, we maximize the function provided in step E with respect to estimated parameters (Equation 7):

$$\hat{\theta} = argmaxQ(\theta) = \text{argmax} \sum_{h=1}^{N}\sum_{jk=1}^{2} T_{jk,h}^{(t)} [\log\tau_j - \log(2\pi) - \frac{1}{2}\log \vee \Sigma_{jk} \vee \frac{1}{2}(x_h - \mu_{jk})^T \Sigma_{jk}^{-1}(x_h\mu_{jk})] \tag{7}$$

The estimated parameters ϕ, (μ_{1k}, Σ_{1k}), and (μ_{2k}, Σ_{2k}) appear in separate terms so that we can independently maximize them. However, for each estimated parameter, the function is maximized by differentiating the likelihood function and setting the derivative as zero. By making appropriate derivations, we obtain the final expression for each parameter of the normal mixture model (Equation 8 and 13).

$$\phi_{1k}^{(t+1)} = \frac{1}{N}\sum_{h=1}^{N}\quad T_{1,jk}^{(t)} \tag{8}$$

$$\phi_{2k}^{(t+1)} = \frac{1}{N}\sum_{h=1}^{N}\quad T_{2,jk}^{(t)} \tag{9}$$

$$\mu_{1k}^{(t+1)} = \frac{\sum_{h=1}^{N}\quad T_{1,jk}^{(t)} x_h^T}{\sum_{h=1}^{N}\quad T_{1,jk}^{(t)}} \tag{10}$$

$$\mu_{2k}^{(t+1)} = \frac{\sum_{h=1}^{N}\quad T_{2,jk}^{(t)} x_h^T}{\sum_{h=1}^{N}\quad T_{2,jk}^{(t)}} \tag{11}$$

$$k\Sigma_1^{(t+1)} \quad = \frac{\sum_{h=1}^{N} T_{1,jk}^{(t)}(x_h - \mu_{1k}^{(t+1)})(x_h - \mu_{1k}^{(t+1)})^T}{\sum_{h=1}^{N}\quad T_{1,jk}^{(t)}} \tag{12}$$

$$\Sigma_{2k}^{(t+1)} \quad = \frac{\sum_{h=1}^{N} T_{2,jk}^{(t)}(x_h - \mu_{2k}^{(t+1)})(x_h - \mu_{2k}^{(t+1)})^T}{\sum_{h=1}^{N}\quad T_{2,jk}^{(t)}} \tag{13}$$

To obtain the results, the packages mixtools (Benaglia, Chauveau, Hunter, & Young, 2010), mvtnorm (Genz et al., 2020), and MASS (Venables & Ripley, 2002) were used in the R software (R Core Team, 2020).

Results and discussion

Because the sensory panel comprised individuals with characteristics described in the methodology, there were two heterogeneous populations that may show differences in terms of the discrimination of specialty coffees. It is assumed that the overall score, which determines the overall classification of coffee as specialty, is influenced by the main sensory attributes body, acidity, and sweetness. Thus, a bivariate sample between these attributes was characterized. In this context, the Gaussian mixture is defined by the combination of these samples in the two populations, trained (G1) and untrained (G2) tasters.

From this perspective, given the recommendations of Redner and Walker (1984), in case the populations of this mixture are poorly separated, it can be expected that the application of the EM algorithm produces few interactions and therefore is efficient for use in the problem addressed in this study, which analyzes the results obtained for proportions of the groups, estimated by the mixture of normal distributions, with responses of each group G1 and G2 considering the covariance structure between the score and the other sensory attributes.

The results in Table 2 showed that the estimation of proportion of individuals belonging to each group is accurate based on the comparison with exact proportions, given by ϕ_1 and ϕ_2. This result indicates that the EM algorithm can be used to distinguish consumer groups and can be used as an inferential procedure for assessing the discrimination of groups of individuals in the composition of a sensory panel applied to the specialty coffee quality analysis.

When the assumption of bivariate normal distribution assumed for sensory was tested, a clear discrimination of the groups was found, as shown by the surface generated by the Gaussian density (Figures 2, 3 and 4) when considering these attributes, which characterizes a bimodal behavior, thus suggesting the discrimination of groups.

Specifically, to confirm the hypothesis that the overall score is influenced by the evaluation of one of the attributes, for each type of coffee, the covariance matrices were estimated for each group, as described in Table 3.

Table 2. Proportions of consumer groups G1 and G2 estimated by the EM algorithm compared to original proportions $\phi_1 = 0.43$, $\phi_2 = 0.57$ considering the covariance between the score and the other attributes.

Coffee	Score and Body	Score and Acidity	Score and Sweetness			
	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
A Error	0.437 ± 0.007	0.563	0.387 ± 0.043	0.613	0.471 0.041	0.529
B Error	0.429 ± 0.001	0.571	0.369 ± 0.061	0.630	0.496 0.066	0.504
C Error	0.398 ± 0.032	0.602	0.414 ± 0.016	0.586	0.422 ± 0.008	0.577
D Error	0.407 ± 0.023	0.593	0.446 ± 0.016	0.554	0.351 ± 0.070	0.649

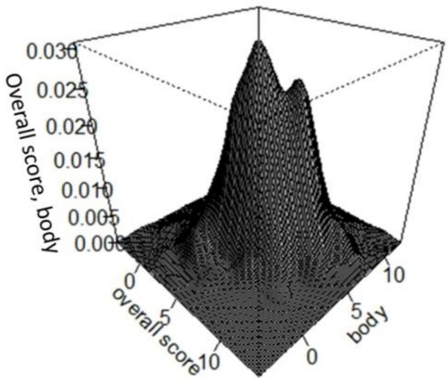


Figure 2. Density of the mixture of normals considering the covariance given by the overall score and body.

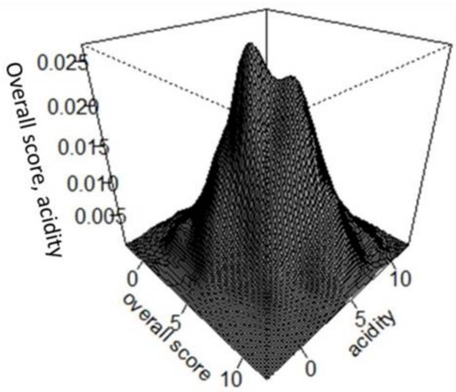


Figure 3. Density of the mixture of normals considering the covariance given by the overall score and acidity.

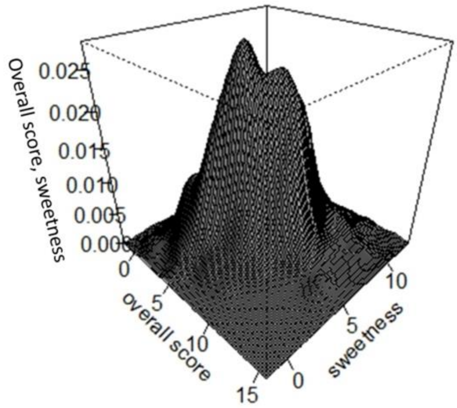


Figure 4. Density of the mixture of normals considering the covariance given by the overall score and sweetness.

Table 3. Estimates of the correlation matrix considering the sensory attribute responses correlated to the score for each type of coffee.

Group (G1), who received training			
Coffee	Score and Body	Score and Acidity	Score and Sweetness
A	$\begin{bmatrix}1 & 0.603\\0.603 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.500\\0.500 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.651\\0.651 & 1\end{bmatrix}$
B	$\begin{bmatrix}1 & 0.625\\0.625 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.596\\0.596 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.775\\0.775 & 1\end{bmatrix}$
C	$\begin{bmatrix}1 & 0.520\\0.520 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.284\\0.284 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.600\\0.600 & 1\end{bmatrix}$
D	$\begin{bmatrix}1 & 0.729\\0.729 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.331\\0.331 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.680\\0.680 & 1\end{bmatrix}$
Group (G2), who did not receive training			
Coffe	Score and Body	Score and Acidity	Score and Sweetness
A	$\begin{bmatrix}1 & 0.381\\0.381 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.156\\0.156 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.545\\0.545 & 1\end{bmatrix}$
B	$\begin{bmatrix}1 & 0.504\\0.504 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.267\\0.267 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.662\\0.662 & 1\end{bmatrix}$
C	$\begin{bmatrix}1 & 0.132\\0.132 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.455\\0.455 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.545\\0.545 & 1\end{bmatrix}$
D	$\begin{bmatrix}1 & 0.476\\0.476 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.205\\0.205 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.542\\0.542 & 1\end{bmatrix}$

The results in Table 3 indicate moderate or low correlations, so there was no evidence that the responses for the sensory attributes acidity, body, and sweetness influenced the overall score, justifying the results reported by Guimarães, Leme, Rezende, Pereira, and Santos (2019) and Santos, Cirillo, and Guimarães (2021) when stating that for consumers less involved with the practice of drinking specialty coffee, their motivations for consumption are essentially based on the taste and aroma of the beverage.

This result raises a discussion about the factors that can justify the poorer discrimination of type C coffee, given the lower estimated correlations. In this context, according to Borém et al. (2020), the sensory quality of coffee at higher altitudes is noticeable. Therefore, comparing the results observed in the correlation of scores with the other attributes, especially the coffee coded in C, characterized by having been produced in a region at lower altitude, compared to the others, there was statistical evidence that justifies the low correlations estimated.

As for the other attributes, a specialty coffee is characterized by sweetness, acidity, and body perceptible to the palate. The intensity and quality of these attributes and the balance between them are used to classify them as higher or lower quality. Thus, coffee samples with high sweetness, high and pleasant acidity, and dense body are, in general, classified as high quality.

On the other hand, coffee samples with low sweetness, unpleasant and pungent acidity, and watery body receive lower scores. While the fragrance and aroma of coffee bring complexity to the beverage, sweetness, acidity, and body are the basic attributes that make up the structure of a coffee beverage and can influence consumer preference.

Pereira et al. (2017) carried out a study with trained tasters with Q-Grader certificates. The methodology followed the analysis protocol guidelines of the Specialty Coffee Association of America, SCAA, with the participation of two testing groups and using Pearson correlation coefficients between fragrance/aroma, uniformity, absence of defects, sweetness, flavor, acidity, body, aftertaste, balance, overall and overall evaluation of the coffees with sensory analyses carried out in the morning and found correlations similar to the results obtained herein.

Regarding the effect of the groups of trained and untrained tasters, for the same specialty coffee, with the attributes of the overall score, aroma, body, and hardness evaluated in a univariate manner as reference and using the distribution of extreme values, Ferreira et al. (2016), considering a significant score worthy of international competition (higher than 9.5), found that the probability of a consumer assigning a score higher than 9.5 is relatively low for all evaluated coffees.

Given the above, the application of the EM algorithm considering bivariate analysis is promising since the correlation between the evaluated attributes and the overall score shows the discrimination of heterogeneous groups, including the estimation of the proportion of the number of individuals belonging to each group.

With the purpose of measuring hit rates regarding the discrimination of specialty coffees, with a sensory panel consisting of groups of trained and untrained tasters, Liska et al. (2015) conducted a study considering the classifiers of the discriminant analysis via a boosting algorithm, in which the training set consisted of 70% original samples and the remaining sample comprised the testing set. Using a comparative method, the authors concluded that the classifier generated by Fisher's discriminant analysis had a reasonable discriminatory power, revealing a high power of discrimination for trained tasters and a low power of discrimination for untrained tasters. Thus, it was observed that the application of the boosting method provided a classifier with a high discriminatory power for the tasters according to their training.

Hit rates were similar to the results reported by Barbosa et al. (2014), in which the discriminant analysis method was adopted to distinguish types of processing of specialty coffees considering different stable isotopes (physiological characteristics) in coffee seeds.

Regarding the performance of the boosting method when extended to more than two groups, Oliveira et al. (2019), for the same database, improved the discriminant analysis by restratifying the groups: T1: trained tasters aged between 19 and 50 years; T2: untrained testers aged between 19 and 50 years; and T3: tasters with experience, but without training, aged over 50 years.

Thus, through the application of the boosting method along with the bagging method (Breiman, 1996), which is a resampling procedure, the authors concluded that the classification error rates were lower than those of the conventional discriminant analysis given by Fisher's linear and quadratic models. In this sense, combination of the multiclass boosting and bagging methods was efficient in capturing small differences in the samples of specialty coffees evaluated (Table 1), indicating their feasibility in the application of results of sensory tests of this nature.

In another experiment with larger groups, Ahmad, Reid, Paulsen, and Sinclair (1999) performed automatic separation of coffee beans. The problem was the differentiation of four types of grains, three of which were characterized by size, whereas the fourth, by size or color.

The classification was performed using image processing algorithms for object detection and feature extraction, classification by color standards, and definition of shape descriptors. For this analysis, the multiclass boosting method (Oliveira et al., 2019) showed error rates lower than 21%.

Conclusion

This case study showed that the EM algorithm can be recommended as a procedure for assessing the proportion of individuals composing a sensory panel, since the estimates were accurate and precise in relation to discrimination of groups with different sensory perceptions.

Estimates of the correlation matrices obtained by the EM algorithm showed that the final score has a greater influence on sweetness, for the four types of coffee evaluated.

In situations where the number of individuals in each group is unknown, the EM algorithm was accurate in estimating the proportion of individuals belonging to each group.

Acknowledgements

The authors would like to thank the National Council for Scientific and Technological Development (*Conselho Nacional de Desenvolvimento Científico and Tecnológico* – CNPq) for the financial support through research in projects 140242/2019-8 and 304939/2021-8.

References

Ahmad, I. S., Reid, J. F., Paulsen, M. R., & Sinclair, J. B. (1999). Color classifier for symptomatic soybean seeds using Image processing. *Plant Disease, 83*(4), 320-327. DOI: https://doi.org/10.1094/PDIS.1999.83.4.320

Barbosa, J. N., Borém, F. M., Alves, H. M. R., Cirillo, M. Â., Sartoli, M. M. P., & Ducatti, C. (2014). Discrimination of production environments of specialty coffees by means of stable isotopes and discriminant model. *Journal of Agricultural Science, 6*(5), 55-64. DOI: https://doi.org/10.5539/jas.v6n5p55

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. S. (2010). Mixtools: an R package for analyzing mixture models. *Journal of Statistical Software, 32*(6), 1-29. DOI: https://doi.org/10.18637/jss.v032.i06

Borém, F. M., Cirillo, M. Â., Alves, A. P. C., Santos, C. M., Liska, G. R., Ramos, M. F., & Lima, R. R. (2020). Coffee sensory quality study based on spatial distribution in the Mantiqueira mountain region of Brazil. *Journal of Sensory Studies, 35*(2), e12552. DOI: https://doi.org/10.1111/joss.12552

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140. DOI: https://doi.org/10.1007/BF00058655

Brookes, C. (2014). Consumer satisfaction: advancements in theory, modeling, and empirical findings. *International Journal of Market Research, 56*(1), 129-131. DOI: https://doi.org/10.2501/IJMR-2014-006

Ceribeli, H. B. (2015). Satisfação do consumidor: percorrendo diferentes perspectivas teóricas. *REUNA, 20*(2), 5-22.

Cirillo, M. Â., Ramos, M. F., Borém, F. M., Miranda, F. M., Ribeiro, D. E., & Menezes, F. S. (2019). Statistical procedure for the composition of a sensory panel of blends of coffee with different qualities using the distribution of the extremes of the highest scores. *Acta Scientiarum. Agronomy, 41*(1), e39323. DOI: https://doi.org/10.4025/actasciagron.v41i1.39323

Dai, H., Luo, X. R., Liao, Q., & Cao, M. (2015). Explaining consumer satisfaction of services: the role of innovativeness and emotion in an electronic mediated environment. *Decision Support Systems, 70*, 97-106. DOI: https://doi.org/10.1016/j.dss.2014.12.003

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, 39*(1), 1-38.

Ferreira, H. A., Liska, G. R., Cirillo, M. Â., Borém, F. M., Ribeiro, D. E., Cortez, R. M., & Guiraldeli, C. H. C. (2016). Selecting a probabilistic model applied to the sensory analysis of specialty coffees performed with consumer. *IEEE Latin America Transactions, 14*(3), 1507-1512. DOI: https://doi.org/10.1109/TLA.2016.7459642

Gashgari, S. (2016). Equity theory and its effect on performance outcome. *International Journal of Scientific & Engineering Research, 7*(4), 517-520.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., … Hothorn, T. (2020). *mvtnorm: multivariate normal and t distributions*. Retrieved from http://mvtnorm.r-forge.r-project.org/

Guimarães, E. R., Leme, P. H. M. V., Rezende, D. C., Pereira, S. P., & Santos, A. C. (2019). The brand new Brazilian specialty coffee market. *Journal of Food Products Marketing,* 25(1), 49-71. DOI: https://doi.org/10.1080/10454446.2018.1478757

Lankton, N. K., & McKnight, H. D. (2012). Examining two expectation disconfirmation theory models: assimilation and asymmetry effects. *Journal of the Association for Information Systems, 13*(2), 88-115. DOI: https://doi.org/10.17705/1jais.00285

Li, Y., Zhang, J., He, R., Tian, L., & Wei, H. (2019). Hybrid DE-EM algorithm for gaussian mixture model-based wireless channel multipath clustering. *International Journal of Antennas and Propagation, 2019*, 4639612. DOI: https://doi.org/10.1155/2019/4639612

Liska, G. R., Menezes, F. S., Cirillo, M. Â., Borém, F. M., Cortez, R. M., & Ribeiro, D. E. (2015). Evaluation of sensory panels of consumers of specialty coffee beverages using the boosting method in discriminant analysis. *Semina: Ciências Agrárias, 36*(6), 3671-3680. DOI: https://doi.org/10.5433/1679-0359.2015v36n6p3671

Makarem, S. C., & Jae, H. (2016). Consumer boycott behavior: an exploratory analysis of twitter feeds. *Journal of Consumer Affairs, 50*(1),193-223. DOI: https://doi.org/10.1111/joca.12080

Oliveira, L. M., Menezes, F. S., Cirillo, M. Â., Saúde, A. V., Borém, F. M., & Liska, G. R. (2019). Machine Learning techniques in muliclass problems with application in sensorial analysis. *Concurrency and Computation*, *32*(7), e5579. DOI: https://doi.org/10.1002/cpe.5579

Oliver, R. L. (2014). *Satisfaction a behavioral perspective on the consumer* (2nd ed.). New York, NY: Taylor & Francis Group.

Ossani, P. C., Cirillo, M. Â., Borém, F. M., Ribeiro, D. E., & Cortez, R. M. (2017). Qualidade de cafés especiais: uma avaliação sensorial feita com consumidores utilizando a técnica MFACT. *Revista Ciência Agronômica, 48*(1), 92-100. DOI: https://doi.org/10.5935/1806-6690.20170010

Pereira, L. L., Cardoso, W. S., Guarçoni, R. C., Fonseca, A. F. A., Moreira, T. R., & ten Caten, C. S. (2017). The consistency in the sensory analysis of coffees using Q-graders. *European Food Research and Technology, 243*, 1545-1554. DOI: https://doi.org/10.1007/s00217-017-2863-9

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, AT: R Foundation for Statistical Computing.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *Society for Industrial and Applied Mathematics, 26*(2), 195-239. DOI: https://doi.org/10.1137/1026034

Santos, P. M., Cirillo, M. Â., & Guimarães, E. R. (2021). Specialty coffee in Brazil: transition among consumers' constructs using structural equation modeling. *British Food Journal, 123*(5), 1913-1930. DOI: https://doi.org/10.1108/BFJ-06-2020-0537

Sørensen, B., Andersen, C., & Andersen, M. P. (2014). Consumer satisfaction and confirmation of habits of comprehension: the effect of inductive print advertisements – Peircean comments. *Semiotica*, *199*, 159-174. DOI: https://doi.org/10.1515/sem-2013-0123

Specialty Coffee Association of America [SCAA]. (2009). *Cupping specialty coffee*. SCAA. Retrieved from https://www.scaa.org/PDF/PR%20-%20CUPPING%20PROTOCOLS%20V.21NOV2009A.pdf

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. (4th ed.). New York, NY: Springer.

Yeretzian, C., Glöss, A., Petrozzi, S., D'ambrosio, L., Knöpfli-Lengweiler, K., Wieland, F., … Anliker, R. (2010). The smell of coffee – an analytical perspective. *New Food*, *13*(4), 10-17.